Εργασία Α3.

Σκοπός της Εργασίας

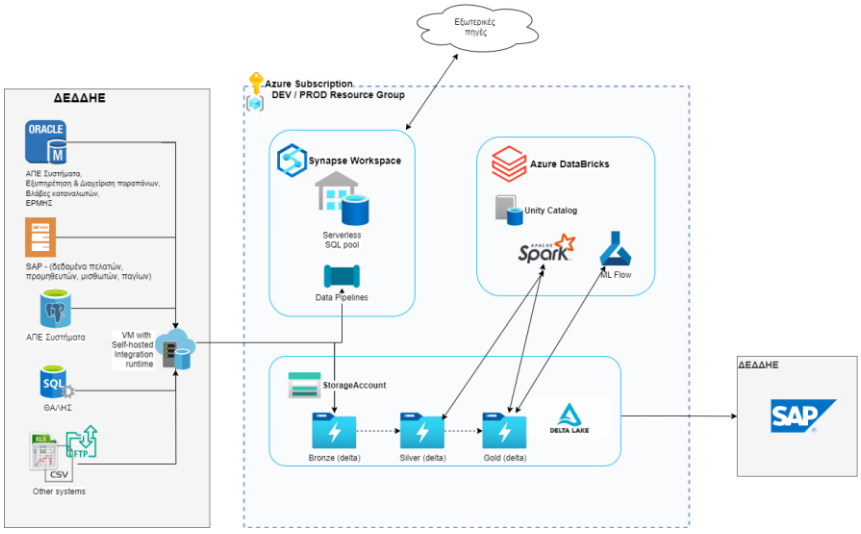
Η παρούσα εργασία είναι μέρος του Παραδοτέου Π1 της Ροής Α και περιγράφει αναλυτικά την πλήρη αρχιτεκτονική του συστήματος και δεδομένων καθώς και τις διεπαφές του με τις υπάρχοντες υποδομές του οργανισμού.

1. System Architecture

Η προτεινόμενη λύση αξιοποιεί τις δυνατότητες που προσφέρονται από τους παρόχους υπολογιστικού νέφους (Cloud providers), και συγκεκριμένα τη δυνατότητα για υλοποίηση Big Data & AI/ML λύσεων σε παραγωγικό περιβάλλον. Στις επόμενες υποενότητες, παρουσιάζεται αναλυτικά η αρχιτεκτονική της προτεινόμενης λύσης, οι τεχνολογίες και τα cloud-native components που θα χρησιμοποιηθούν.

Ακολουθεί η παρουσίαση της λειτουργικής αρχιτεκτονικής (Functional Architecture) της προτεινόμενης λύσης.

Application architecture



Azure Synapse Workspace: Η συγκεκριμένη υπηρεσία θα χρησιμοποιηθεί σαν μια κεντρική αποθήκη δεδομένων αφού οι επεξεργασίες που έχουν γίνει σε προηγούμενα

στάδια (Azure Data Factory και Azure Databricks) μεταφέρονται στην τελική τους μορφή (ζώνη Gold) σε μια βάση δεδομένων στο Azure Synapse. Αποτελεί την βάση για όλες τις μετέπειτα διαδικασίες προετοιμασίας αναφορών (reporting) και αναλύσεων δεδομένων. Είναι μια ολοκληρωμένη πλατφόρμα επεξεργασίας και ανάλυσης δεδομένων που ενσωματώνει λειτουργίες και υπηρεσίες για την αποθήκευση, την επεξεργασία και την ενοποίηση μεγάλου όγκου δεδομένων. Οι υπηρεσίες αυτές περιλαμβάνουν SQL servers είτε σε dedicated ή serverless pools ενώ πάνω στα δεδομένα μπορεί να έχει επίσης πρόσβαση ένα σμήνος Spark workers για την επεξεργασία τους. Αυτή η υπηρεσία αποτελεί το κέντρο του αποθετηρίου του "Gold" επιπέδου δεδομένων και θα υποστηρίζεται με ένα dedicated sql pool για τη κατανάλωση τους από συστήματα που εξυπηρετεί. Το Azure Synapse αποτελεί τη κύρια υποδομή Big Data, καθώς όλοι οι πίνακες ακολουθούν μια λογική πολλαπλής κατανομής (multi distributed) προκειμένου η προσπέλαση και χρήση των δεδομένων να είναι η βέλτιστη σε κάθε αναζήτηση (συμπίεση, ταχύτητα απόκρισης). Ιδιαίτερα χαρακτηριστικά τις υπηρεσίας:

α. Παράλληλη επεξεργασία σε κάθε query μέσω κατανεμημένων nodes (distributed)
β. Αυτοματοποιημένα ημερήσια backups σε γεωγραφική εφεδρεία (geo redundant)
γ. Υψηλή διαθεσιμότητα μέσω κατάλληλης παραμετροποίησης (failover group)
δ. Εγγενή server-side κρυπτογράφηση δεδομένων (encryption at rest, transparent data encryption TDE), TLS κρυπτογράφηση κατά την μεταφορά (encryption in transit)

Για λόγους ασφάλειας, η κάθε εφαρμογή χρησιμοποιεί την υπηρεσία Azure Key Vault για την αποθήκευση, χρήση και διαχείριση των κλειδιών ή ευαίσθητων δεδομένων χρειάζονται για τη λειτουργία της (κωδικοί βάσεων δεδομένων κτλ.).

Όλη η ανωτέρω αναφερόμενη υποδομή του αρχιτεκτονικού διαγράμματος, αποτελεί τη βάση για τη δημιουργία τριών διαφορετικών αλλά πανομοιότυπων περιβαλλόντων:

1. **STAGING/UAT:** Περιβάλλον δοκιμών, ποιοτικού ελέγχου και εκπαίδευσης
2. **PROD:** Περιβάλλον παραγωγικής λειτουργίας του συστήματος

Commented [FS1]: 2 envs

Ο σχεδιασμός της υποδομής έχει προβλέψει τις παρακάτω λογικές οριοθετήσεις των επιμέρους υπηρεσιών Azure από τους οποίους υπάρχει ροή πληροφορίας ή/και πρόσβασης στο συνολικό σύστημα.

Στο αρχιτεκτονικό διάγραμμα συμπεριλαμβάνει τα εξής:

Πηγές Δεδομένων ΔΕΔΔΗΕ

Η λύση τροφοδοτείται από μια σειρά εξωτερικών και εσωτερικών πηγών (δηλαδή πηγών που δεν ανήκουν ή ανήκουν στον ΔΕΔΔΗΕ αντίστοιχα). Θα υποστηρίζεται η διαλειτουργικότητα με διάφορες πηγές δεδομένων), όπως:

- Σχεσιακές βάσεις είτε από τα συστήματα on-premises του ΔΕΔΔΗΕ, είτε περιβάλλοντα υπολογιστικού νέφους (Ενδεικτικά και σε σχέση με τα δεδομένα της διακήρυξης Oracle, SQL Server, PostgreSQL, DB2, κ.ά., ενώ θα γίνουν προσαρμογές αν χρειαστεί και για άλλες βάσεις δεδομένων).
- Data lakes, ενδεικτικά και όχι περιοριστικά, Hadoop Distributed File System (HDFS), Azure Data Lake.
- Μεμονωμένα αρχεία, ενδεικτικά τύπου CSV, excel, text, JSON που μπορούν να ανακτώνται:
    α) με πρωτόκολλα μεταφοράς όπως *FTP (File Transfer Protocol)*, *Secure File Transfer Protocol (SFTP)*, και
    β) μέσω συστημάτων διαχείρισης αρχείων (*Document Management Systems - DMS).* Ενδεικτικά, τέτοιες λύσεις διαμοιρασμού αρχείων όπως τα Microsoft OneDrive, Sharepoint.

Από τα παραπάνω, γίνεται κατανοητό ότι η λύση έχει τη δυνατότητα να αποθηκεύσει και να διαχειριστεί δομημένα, μη-δομημένα & ημι-δομημένα δεδομένα (structured, unstructured και semi-structured) εφόσον χρειαστεί. Να σημειωθεί, ωστόσο ότι στη διακήρυξη διαφαίνεται ότι θα είναι δομημένα και ημιδομημένα.

Σχετικά με τις εξωτερικές πηγές, το σύστημα εξασφαλίζει διαλειτουργικότητα με αυτές είτε μέσω των κατάλληλων web services/ REST APIs, είτε μέσω FTP, SFTP, DMS όπως αναφέρθηκε παραπάνω. Η λύση επιτρέπει την ανάκτηση δεδομένων από διάφορες πηγές και με διαφορετικά πρωτόκολλα επικοινωνίας όπως αυτές θα

οριστούν κατά την ανάλυση και καταγραφή εργασιών των απαιτήσεων διασύνδεσης με εξωτερικές πηγές πληροφόρησης

Self-hosted Integration Runtime

Για τη μεταφορά δεδομένων απο τις υποδομές εσωτερικού δικτύου (on-prem) θα εγκατασταθεί ενδιάμεση υπηρεσία (Self-hosted Integration Runtime) σε τοπικό υπολογιστή εντός του εσωτερικού δικτύου του ΔΕΔΔΗΕ, που διαθέτει πρόσβαση τόσο τις τοπικές πηγές δεδομένων όσο και στο Azure. Η υπηρεσία αυτή, παρέχει μια αποκλειστική γέφυρα συνδεσιμότητας μεταξύ των τοπικών βάσεων δεδομένων και του Synapse που χρειάζεται για να απορροφήσει δεδομένα και επιτρέπει μια υβριδική ενσωμάτωση δεδομένων στη νέα υποδομή cloud χωρίς να απαιτείται να δημιουργούνται ειδικές εισερχόμενες θύρες δικτύου. Προκειμένου η πρόσβαση στα δεδομένα να είναι εφικτή, ο ενδιάμεσος υπολογιστής (VM) που φιλοξενεί την υπηρεσία θα είναι κατάλληλα παραμετροποιημένος (Oracle Data Access Components (ODAC), Oracle client, SQL Developer κοκ.) και η σύνδεση με τις εκάστοτε βάσεις δεδομένων θα είναι επαληθευμένη, δηλ. η σύνδεση με τις βάσεις να λειτουργεί. Η διασύνδεση του on premise VM με το Data Factory στο Azure προτείνεται να γίνεται μέσω Azure Private Link που θα διατεθεί από τη ΓΓΠΣ το οποίο δίνει τη δυνατότητα πρόσβασης σε υπηρεσίες Azure PaaS μέσω ενός ιδιωτικού τερματικού σημείου (private endpoint) στο εικονικό δίκτυο της πηγές δεδομένων. Η κίνηση μεταξύ του εικονικού δικτύου και της υπηρεσίας ταξιδεύει στο βασικό δίκτυο (backbone) της Microsoft και ως εκ τούτου η ταχύτητα είναι πολύ μεγάλη ενώ έκθεση της υπηρεσίας και των δεδομένων στο δημόσιο διαδίκτυο αποφεύγεται.

Για τις περιπτώσεις σύνδεσης συστημάτων που βρίσκονται εντός του Azure, γίνεται μέσω το private endpoint για κάθε σύστημα/υπηρεσία στο εικονικό δίκτυο (virtual network) που φιλοξενεί την πηγή των δεδομένων.

Web Application Firewall (WAF): Πρόκειται για υπηρεσία προστασίας η οποία, για λόγους ασφάλειας, δρομολογεί την κίνηση σε ένα συγκεκριμένο endpoint (η επικοινωνία αυτή είναι κρυπτογραφημένη με TLS πρωτόκολλο). Παρέχει τη δυνατότητα να τερματίζει όποιες συνδέσεις θεωρηθούν ως πιθανές απόπειρες μη εξουσιοδοτημένης πρόσβασης ή/και επιθέσεις.

Virtual Network (Vnet): Πρόκειται για υπηρεσία δημιουργίας ιδιωτικού δικτύου εντός του Azure cloud απομονωμένο από το δημόσιο Internet ή άλλα εικονικά δίκτυα. Για λόγους ασφάλειας, το Vnet δεν επικοινωνεί με άλλα δίκτυα ή υπηρεσίες στο Azure εκτός και αν γίνει ειδική διασύνδεση μεταξύ αυτών (Vnet peering, private endpoints)

Private Endpoint: Πρόκειται για υπηρεσία όπου παρέχει πρόσβαση εγκεκριμένων χρηστών ή εφαρμογών σε άλλες Azure υπηρεσίες τοποθετημένων εντός εικονικών δίκτυών διαμέσου αποκλειστικής σύζευξης

Azure Data Lake Storage Gen2: Σε αυτή την υπηρεσία αποθηκεύονται κάθε είδους δεδομένα, είτε πρωτογενή προερχόμενα από τα πηγαία συστήματα χωρίς κάποιου είδους προσαρμογή ή τα αποτελέσματα ανάλυσης που παράγονται από τα διάφορα στάδια και ροές επεξεργασίας τα οποία τοποθετούνται σε ειδικά διαμορφωμένους φακέλους (containers) ακολουθώντας μια ιεραρχική δομή αποθήκευσης σε διαφορετικά επίπεδα (zones) ακολουθώντας τα πρότυπα αρχιτεκτονικής Delta lake

Ιδιαίτερα χαρακτηριστικά τις υπηρεσίας:

α. Ανοχή σφαλμάτων (fault-tolerance)
β. Επεκτασιμότητα χωρίς όριο (scalability)
γ. Εισαγωγή μεγάλου όγκου δεδομένων με υψηλό ρυθμό διαμεταγωγής (high-throughput)
δ. Εισαγωγή μεγάλου όγκου αρχείων μικρού μεγέθους (small writes) με πολύ μικρή καθυστέρηση (low latency)
ε. Κρυπτογράφηση δεδομένων (Encryption at Rest)
στ. Πολλαπλά αντίγραφα σε γεωγραφική εφεδρεία (geo-redundant)
ζ. Υψηλή διαθεσιμότητα εξ ορισμού (24x7)
η. Ασφαλή σύνδεση μέσω αποκλειστικής σύζευξης (private endpoint)

Ένα επιπλέον επίπεδο οργάνωσης των δεδομένων σε αυτό το μέσο, διαμορφώνεται ανάλογα με τις ιδιαίτερες ανάγκες διακράτησης της ιστορικότητας (ως προς το χρονικό εύρος) αλλά και της αμεσότητας στην πρόσβαση της πληροφορίας που απαιτείται ανά επιμέρους εφαρμογή και το οποίο αναλύεται ως εξής:

- **«Θερμό»** Επίπεδο Πρόσβασης (Hot Access Tier): Τα δεδομένα είναι προσβάσιμα σε πραγματικό χρόνο και χρησιμοποιούνται για συχνές και συνεχείς λειτουργίες. Έχει υψηλότερο κόστος αποθήκευσης σε σχέση με τα άλλα δύο επίπεδα και η περίοδος διακράτησης είναι μέχρι τα 5 έτη

- **«Ψυχρό»** Επίπεδο Πρόσβασης (Cold Access Tier): Αποθηκεύονται τα δεδομένα που δεν απαιτούν συχνή χρήση ή τροποποίηση και είναι πιο οικονομική από το προηγούμενο επίπεδο. Η περίοδος διακράτησης είναι μεταξύ των 5 και 10 ετών
- Επίπεδο Πρόσβασης **Αρχειοθήκης** (Archive Access Tier): Αφορά δεδομένα των οποίων η χρήση απαιτείται (πολύ) σπάνια, και η πρόσβαση είναι πολύ περιορισμένη και πάντως όχι άμεση (σε επίπεδο ωρών). Τα δεδομένα σε αυτό το επίπεδο πρακτικά βρίσκονται εκτός σύνδεσης (offline) και, προκειμένου να αξιοποιηθούν, πρέπει να έχουν μετακινηθεί σε ένα από τα άλλα δύο επίπεδα ώστε να αποκατασταθεί η πρόσβαση σε αυτά. Η αποθήκευση είναι η πιο οικονομική περίπτωση ως προς το κόστος διακράτησης, το κόστος πρόσβασης όμως είναι ιδιαίτερα υψηλό. Η περίοδος διακράτησης είναι άνω των 10 ετών

Azure Databricks: Αποτελεί τη βασική πλατφόρμα, επεξεργασίας δεδομένων και μηχανικής μάθησης, όπως επίσης και βασικό εργαλείο δημιουργίας του Delta Lake. Το Delta Lake είναι μια υπηρεσία αποθήκευσης του Azure (Storage Account) το οποίο υποστηρίζει ιεραρχική δομή δεδομένων. Το Azure Databricks συνδέεται με το Azure storage (mount) και μπορεί να επεξεργαστεί τα δεδομένα εκεί. Τα δεδομένα αποθηκεύονται στο Delta Lake, είτε σε αρχεία parquet ή delta parquet -ανάλογα με το στάδιο (bronze/silver/gold ζώνη) που προορίζεται να καταλήξουν- ακολουθώντας μια ιεραρχία στη δημιουργία φακέλων και αρχείων (taxonomy). Θα αξιοποιηθεί επίσης η δυνατότητα του Unity Catalog για δυνατότητες κεντρικού ελέγχου πρόσβασης, επιθεώρησης, γενεαλογίας και ανακάλυψης δεδομένων στο Azure Databricks Ιδιαίτερα χαρακτηριστικά τις υπηρεσίας:

α. Ανοχή σφαλμάτων (fault-tolerance) και μηδενική απώλεια δεδομένων μέσω εγγραφών σε ειδικά logs (Write Ahead Logs)
β. Πολλαπλά αντίγραφα σε γεωγραφική εφεδρεία (geo-redundant) για το περιβάλλον Ελέγχου και Δεδομένων (workspace, VM clusters, storage accounts)
γ. Κρυπτογράφηση δεδομένων με κλειδιά πελάτη (customer-managed keys) για όλο το περιβάλλον ή μέρη αυτού
δ. Υψηλή διαθεσιμότητα εξ ορισμού (24x7)

Η μετάπτωση των δεδομένων, ο μετασχηματισμός τους και η περαιτέρω επεξεργασία τους γίνεται μέσω του οικοσυστήματος ***Databricks***. Τα ***Databricks Notebooks*** τα οποίο σε κώδικα *PySpark (βασισμένο σε Python)* θα υλοποιούν όλες τις διεργασίες (Jobs) επεξεργασίας και ανάλυσης των δεδομένων. Υπάρχει δε δυνατότητα χρονοπρογραμματισμού και αυτοματοποίησης αυτών των διεργασιών.

Το οικοσύστημα αυτό θα μας δώσει μια σειρά δυνατοτήτων που θα καλύπτουν functional και non-functional προαπαιτούμενα. Συγκεκριμένα, θα υποστηρίξει

1. Τις διαδικασίες που εκτελούνται από τους μηχανικούς Βάσεων Δεδομένων (data engineering εργασίες) γύρω από τον άξονα εξαγωγή, μετασχηματισμός, φόρτωση (Extract, Transform, Load – ETL).
2. Τον καθαρισμό, εμπλουτισμό και την ανάπτυξη αλγορίθμων μηχανικής μάθησης και τεχνητής νοημοσύνης με χρήση διαφόρων βιβλιοθηκών ανοικτού κώδικα.
3. Την κατανεμημένη επεξεργασία δεδομένων που είναι απαραίτητη για μεγάλου όγκου δεδομένα (Big Data).
4. Την συνολική διαχείριση/ενορχήστρωση της λύσης, με δυνατότητες monitoring και logging.

Για τη συνολική λύση της αποθήκευσης των δεδομένων στις διάφορες φάσεις, από την αρχική μετάπτωση μέχρι τον τελικό μετασχηματισμό που θα καλύπτει τις ανάγκες του νέου ολοκληρωμένου Πληροφοριακού Συστήματος Ηρακλής (SAP), προκρίνεται το cloud-native/ managed δομικό στοιχείο **Azure Data Lake (Gen 2)**, όπου δίνει δυνατότητες αποθήκευσης δεδομένων οποιουδήποτε μεγέθους, σχήματος και ταχύτητας. Οι δυνατότητες της λύσης θα επεκταθούν περαιτέρω με τη χρήση του **Delta Lake**, ενός storage layer σχεδιασμένο και βελτιστοποιημένο να τρέχει πάνω από ένα data lake με σκοπό την βελτίωση της αξιοπιστίας, της ασφάλειας και της απόδοσης. Υποστηρίζει δε επεκτάσιμα μεταδεδομένα, ενοποιημένες ροές και επεξεργασία batch δεδομένων. Καταγράφει όλες τις αλλαγές (Deltas) που γίνονται στα δεδομένα σε ένα σειριακό αρχείο καταγραφής συναλλαγών, προστατεύοντας την ακεραιότητα και την αξιοπιστία των δεδομένων και παρέχοντας πλήρεις, ακριβείς διαδρομές ελέγχου. Συμπληρωματικά, η λύση θα παρέχει δυνατότητες αντιγράφων ασφαλείας με τις

προδιαγραφές που θα οριστούν από τον ΔΕΔΔΗΕ είτε ως αντίγραφο στο ίδιο data lake, είτε σε άλλο ανεξάρτητο data lake/ βάση δεδομένων σύμφωνα με τις υποδείξεις του ΔΕΔΔΗΕ.

Πρέπει τέλος να τονιστεί ότι η λύση αυτή θα αναπτυχθεί σε δύο περιβάλλοντα:

- **Staging (STAGE)**: Είναι ένα δοκιμαστικό περιβάλλον που δοκιμάζονται σε μεγάλη κλίμακα (με μέρος ή και όλα τα δεδομένα) οι διάφορες διεργασίες που αναπτύσσονται στα Databricks Notebooks αλλά και τα υλοποιημένα services/ applications της λύσης πριν την πλήρη ανάπτυξή τους στο παραγωγικό περιβάλλον.
- **Production (PROD)**: Είναι το παραγωγικό περιβάλλον όπου εκεί θα αναπτύσσονται (deployment) οι εφαρμογές/ διεργασίες εφόσον έχει επαληθευτεί η σωστή και εύρυθμη λειτουργία τους και θα επικοινωνεί με τα παραγωγικά περιβάλλοντα των άλλων συστημάτων που θα αλληλεπιδρούν με τη λύση.

Συνοψίζοντας, πρέπει να τονιστεί ότι το σύστημα εξασφαλίζει την ανάπτυξη (deployment), ευρωστία (robustness) και επεκτασιμότητα (scalability) της λύσης σε οποιοδήποτε φόρτο. Προσφέρει δε την επίτευξη της μέγιστης δυνατής διαλειτουργικότητας μεταξύ της λύσης και των εξωτερικών συστημάτων, τη προσαρμογή της λύσης στις ειδικές απαιτήσεις ΔΕΔΔΗΕ που θα προκύψουν κατά κύριο λόγο στην Μελέτη Μεθοδολογίας του έργου και στην Ανάλυση Υφιστάμενης Κατάστασης, Συστημάτων & Δεδομένων (Ροές Α και Β).

Αρχιτεκτονική δεδομένων

Για τη διαχείριση και οργάνωση των δεδομένων θα χρησιμοποιηθεί το Unity Catalog Περισσότερες πληροφορίες παρέχονται στο ακόλουθο link. (https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/). Αποτελεί ένα κεντρικό αποθετήριο που περιέχει πληροφορίες σχετικά με τα

διαθέσιμα σύνολα δεδομένων, τους πίνακες, τις προβολές και άλλα στοιχεία που σχετίζονται με τα δεδομένα.
Το Unity Catalog λειτουργεί ως ένα κεντρικό σημείο διαχείρισης, επιτρέποντας την επεξεργασία, την ανάκτηση και την ανάλυση των δεδομένων. Επιτρέπει επίσης, την οργάνωση των δεδομένων σε διάφορα επίπεδα και κατηγορίες, επιτρέποντας τη διαίρεση των δεδομένων σε λογικά τμήματα για ευκολότερη διαχείριση. Υποστηρίζει επίσης λειτουργίες, όπως τη δημιουργία και τη διαχείριση πινάκων, τη δημιουργία προβολών για την αποτίμηση και τη μετασχηματισμό των δεδομένων, καθώς και τη δυνατότητα εκτέλεσης πολύπλοκων ερωτημάτων για την εξαγωγή πληροφοριών από τα δεδομένα. Παρέχει μια ενιαία πλατφόρμα για την ανάπτυξη, την ανάλυση και την εξόρυξη γνώσης από τα δεδομένα, επιτρέποντας στους χρήστες να αξιοποιήσουν πλήρως τις δυνατότητες του Databricks για να εργαστούν με τα δεδομένα τους με αποδοτικό τρόπο.
Η εμβάθυνση του μοντέλου δεδομένων θα πραγματοποιηθεί στη φάση υλοποίησης και θα καταγραφεί στο τεχνικό documentation του παραδοτέου συστήματος όπου θα είναι δυνατή η δειγματοληψία από όλα τα παραγωγικά δεδομένα των εξωτερικών πηγών και θα οριστικοποιηθούν οι διαδικασίες μηχανικής μάθησης.

Το μοντέλο δεδομένων βασίζεται σε 3 ζώνες / κατηγορίες αντικειμένων:
Η "χάλκινη" ζώνη (bronze) συνήθως αποθηκεύει δεδομένα στη αρχική ανεπεξέργαστη τους μορφή. Περιέχει αφιλτράριστα δεδομένα με τα ακόλουθα χαρακτηριστικά:

1. Ακατέργαστη μορφή όπως έρχεται από την πηγή
2. Γίνονται partitioned ανάλογα με το χρονικό διάστημα παραλαβής ή ημερομηνίας που αναφέρονται.
3. Αποτελούν τα views η exports που έχουν ρυθμιστεί στα πηγαία συστήματα και μεταφέρονται σε batch ή stream μορφή
4. Έχουν περιορισμένο TTL (time to live) και αρχειοθετούνται ή διαγράφονται μετά την επεξεργασία και ενημέρωση του silver layer αφού είναι εφικτή η επαναφορά τους από τα πηγαία συστήματα αν χρειαστεί.

Η ασημένια ζώνη (Silver) παρέχει μια πιο εμπλουτισμένη δομή των δεδομένων που έχουν αντληθεί. Αντιπροσωπεύει μια επικυρωμένη, εμπλουτισμένη έκδοση των δεδομένων που μπορεί να είναι αξιόπιστη για μεταγενέστερες εργασίες, τόσο λειτουργικές όσο και αναλυτικής. Επιπλέον, το Silver zone έχει τα ακόλουθα χαρακτηριστικά:

1. Χρησιμοποιεί κανόνες ποιότητας δεδομένων για την επικύρωση και την επεξεργασία δεδομένων.

2. Συνήθως περιέχει μόνο λειτουργικά δεδομένα. Έτσι, τεχνικά δεδομένα ή άσχετα δεδομένα από το Bronze φιλτράρονται.
3. Γίνεται καθαρισμός των δεδομένων, ελλείπουσες τιμές, κανονικοποίηση κ.α.
4. Τα δεδομένα συνήθως εμπλουτίζονται με δεδομένα αναφοράς.
5. Τα δεδομένα είναι ομαδοποιημένα γύρω από ορισμένες θεματικές περιοχές.
6. Τα δεδομένα εξακολουθούν να είναι οργανωμένα ανά πηγή-σύστημα.
7. Έχουν TTL ισοδύναμο με το Gold layer

Τα δεδομένα από την χρυσή ζώνη (Gold), σύμφωνα με τις αρχές μιας αρχιτεκτονικής Lakehouse, οργανώνονται σε βάσεις δεδομένων έτοιμες για κατανάλωση. Από αυτή την άποψη, η ιδιοκτησία των δεδομένων αλλάζει, επειδή τα δεδομένα δεν είναι πλέον ευθυγραμμισμένα με την πηγή τους. Αντίθετα, έχουν ενσωματωθεί και συνδυαστεί με άλλα δεδομένα του Silver layer ή και του Gold layer.

1. Οι χρυσοί πίνακες αντιπροσωπεύουν δεδομένα που έχουν μετατραπεί για να εξυπηρετήσουν τους τελικούς σκοπούς του συστήματος σε επίπεδο analytics, μηχανικής μάθησης ή παρουσίασης. Τα δεδομένα αποθηκεύονται σε Delta Lake για να είναι δυνατή η ενημέρωσή τους από μελλοντικά δεδομένα αλλά και η σταδιακή (incremental) επεξεργασία μόνο των νέο-εισερχόμενων δεδομένων και όχι του συνόλου κάθε φορά.
2. Υποστηρίζει εκδόσεις και time-travel ερωτήματα για την σύνθετη ανάλυση των δεδομένων σε χρονικά διαστήματα (time window analysis)
3. Στο Gold zone εφαρμόζονται σύνθετοι επιχειρησιακοί κανόνες. Έτσι, χρησιμοποιεί πολλές δραστηριότητες μετα-επεξεργασίας και υπολογισμού για συγκεκριμένες χρήσεις.
4. Τα δεδομένα ελέγχονται σε μεγάλο βαθμό και είναι καλά τεκμηριωμένα.

2. Network Architecture